Explainable Lung Disease Detection from Chest X-Rays with Uncertainty Estimation: A Review

Poonam Waghmare[1], Akshay Hirkar[2], Dhiraj Doijad[3], Pratik Bhingare[4], Prof. D. A. Kamble[5]
[1,2,3,4]*Undergraduate Students, Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Maharashtra, India*
[5]*Asst. Professor, Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Maharashtra, India*

***Abstract*—Deep learning systems can identify many lung diseases from chest X-rays nearly as accurately as radiologists. Those results, however, usually come from just one dataset. Performance often falls when the model encounters images from a different hospital or scanner, or images labelled through another process; worse, it rarely signals that its prediction may be unreliable. This paper examines research addressing that problem through three questions: whether chest X-ray classifiers work across datasets, whether clinicians can examine the reasoning behind their decisions, and whether the systems can communicate their confidence. We review major public datasets, convolutional baselines such as DenseNet121, Grad-CAM explanations, and uncertainty estimation through methods such as Monte Carlo dropout, deep ensembles, and calibration. The discussion also covers our evaluation of TrustXray, including early NIH ChestX-ray14 validation results measured by area under the ROC curve (AUC). Cross-dataset results from CheXpert are still being collected. We close by identifying the open problems that matter most.**

***Index Terms*—Calibration, Chest X-ray, Cross-dataset generalization, Deep learning, Explainable AI, Grad-CAM, Uncertainty estimation.**

I. INTRODUCTION

When a patient arrives coughing, with chest pain or difficulty breathing, the first scan is often a chest X-ray. It is fast, inexpensive, and widely available. Those same advantages create a problem: radiologists in crowded hospitals may need to interpret huge numbers of films. That pressure has driven extensive work on neural networks for chest-image analysis. After large labelled datasets were released [1], [2], convolutional networks matched or surpassed radiologists on certain narrowly defined tasks. CheXNet, designed to detect pneumonia, remains the most familiar example [3].

Headline numbers like these deserve a second look. Most of them are measured on a test set that comes from the same source as the training set. When researchers take a trained model and run it on images from a different hospital, the scores usually fall, sometimes by a lot [4], [5]. A network can pick up on things that have nothing to do with disease, such as the type of scanner, a side marker, or the way a hospital takes portable films, and those habits do not travel [6]. On top of that, a standard classifier gives every image the same kind of confident-looking answer. It never says that it has not seen anything like this before.

These systems are supposed to be more trustworthy thanks to two main ideas. Explanation methods try to highlight which parts of an image led to a prediction so a clinician can see if the model focused on the lungs, not just a label stuck in the corner. Uncertainty estimation gives an honest confidence score for each prediction, so cases the model isn't sure about go straight to a human. Still, neither solution fixes everything, and both have faced criticism. Let's take a fair look at that below.

Most of this is a review, with a bit of experimental work thrown in. Section II dives into the datasets, Section III goes over the baseline models, and Section IV looks at how well things generalize across different datasets. Section V gets into explainability, while Section VI tackles uncertainty. Then, in Section VII, we lay out how we evaluated our own project, TrustXray, and share some initial validation results. Section VIII wraps things up with a list of open problems.

II. DATASETS AND LABELS

A. Publicly Available Chest X-ray Datasets

Several publicly accessible chest X-ray datasets are widely used to develop and evaluate deep learning models for medical image analysis, as summarized in Table I. The NIH ChestX-ray14 dataset provides a large collection of X-ray images and has supported the development of deep learning approaches for identifying thoracic diseases [1]. CheXpert, introduced by researchers at Stanford University, offers another extensive collection and includes uncertain labels to represent cases in which a finding cannot be classified confidently as either present or absent [2]. MIMIC-CXR provides chest radiographs alongside their corresponding clinical reports, allowing researchers to investigate relationships between image features and written medical observations [7]. In contrast, PadChest, developed using data from a Spanish hospital, contains a broad range of radiographic findings and supports research involving a wider variety of diagnostic categories [8]. Together, these datasets provide valuable resources for training and evaluating AI models, while their differences in image sources, annotation practices, and diagnostic labels make cross-dataset evaluation particularly important.

TABLE I Commonly used chest X-ray datasets

Dataset	Images	Patients	How labels were made
NIH ChestX-ray14	112,120	30,805	Text mined from reports, 14 labels
CheXpert	224,316	65,240	Rule-based labeler, 14 observations, with uncertain labels
MIMIC-CXR	377,110	65,379	Labels derived from free-text reports
PadChest	160,000+	67,000+	174 findings, partly radiologist labelled

B. Challenges Associated with Dataset Labels

The reliability of dataset annotations is an important concern when developing AI models for chest X-ray interpretation. Because manually reviewing and labelling a very large number of medical images requires substantial time and expert effort, many datasets use automated techniques to extract diagnostic information from radiology reports. Although this approach makes large-scale data collection possible, the extracted labels may not always accurately represent the abnormalities visible in an image. Previous research has identified inconsistencies between some labels and the corresponding X-ray findings in the NIH dataset [9]. Such errors can affect model training and make it difficult to determine the true diagnostic performance of a system.

Differences in annotation procedures across datasets introduce another challenge, as two collections may use different approaches to assign labels to similar findings. Consequently, a reduction in AUC during cross-dataset testing may reflect differences in image characteristics, annotation quality, or both, rather than changes in image data alone.

III. CONVOLUTIONAL BASELINES

Most people start with DenseNet121—a neural network where each layer gets all the feature maps from the layers before it. It's compact enough to run on a single consumer GPU, but still powerful; a lot of research papers use it as the main baseline. CheXNet takes DenseNet121 and fine-tunes it on the ChestX-ray14 dataset. When researchers first tested it, CheXNet actually outperformed the average score of four radiologists on pneumonia detection.

Chest X-ray labeling is a multi-label task since a patient could show signs of both effusion and atelectasis, you need to track each finding separately. The model outputs one sigmoid score per possible condition, working to minimize the binary cross-entropy for each label and then averages the loss across all C labels. For any specific label c, where yc is the true value and pc is the predicted probability, the loss formula looks like this:

$$\ell_c = -y_c \log p_c - (1 - y_c) \log (1 - p_c) \quad (1)$$

We report results as AUC for each label and then average them. AUC is handy since it isn't tied to any specific threshold, but it can be misleading. If you have a rare finding with just a few positive cases in your test set, the AUC might look okay, but it doesn't really tell you much. That's why you should always show confidence intervals alongside the AUC.

IV. CROSS-DATASET GENERALIZATION

One study used two datasets to train pneumonia models and then tested on a new hospital. It showed good performance on internal testing however performance decreased when testing externally. The network was also able to predict the source of the images. This is bad because hospitals can have differing levels of severity for their patients. Another study used just one public dataset to train and found a similar decrease in accuracy for chest x-ray classification [5]. Another group found similar conclusions on how transferable the models were [11]. In the case of detecting COVID-19 it was found that several models were using shortcuts from the image rather than the lungs [6].

Various remedies have been tried. The cheapest is data augmentation, where random crops, rotations and contrast changes are applied so that the network does not rely on exact pixel statistics [13]. Domain-adversarial training adds a second network that tries to guess which dataset the data came from, and the main network is encouraged to make that guess hard [12]. There are two other options: training on multiple datasets at the same time, and harmonising the image intensities prior to training. None of them have fully closed the gap and all of them add cost or complexity. Testing across datasets is an issue due to the lack of identical labels. There are only 7 overlaps in the labels between the NIH and CheXpert data sets (atelectasis, cardiomegaly, consolidation, edema, pleural effusion, pneumonia and pneumothorax). This means when a model is tested on these it can only be done on those 7 labels. Also, the labels may be worded differently meaning that even a perfect model would not get 100% accuracy.

V. EXPLAINABILITY

Grad-CAM is the go-to method in most chest X-ray papers [14]. The idea is simple. You take the feature maps from the last convolutional layer, A^k, and ask how much each one affects the score y^c for class c. To get the importance weight of map k, you average the gradient of that score over all Z positions in the map:

$$\alpha^c_k = (1/Z)\ \Sigma_{i,j}\ \partial y^c/\partial A^k_{ij} \quad (2)$$

and the heatmap is a weighted sum of the maps, with negative values removed,

$$L^c = \mathrm{ReLU}(\Sigma_k\ \alpha^c_k\ A^k) \quad (3)$$

That small map is then stretched to the size of the image and laid over the X-ray as a heatmap. The catch is resolution. With a DenseNet121 and a 224 by 224 input, the last feature map is only 7 by 7, so the heatmap comes out blurry. It can show that the model was looking at the lower left lung, but it can't pick out something as small as a nodule. It can tell you that the model looked at the lower left lung. It cannot outline a small nodule.

Grad-CAM has some competitors. LIME builds a small, simple model around one prediction to explain it [15]. SHAP splits the prediction into shares and gives each region its part [16]. Integrated gradients start from a blank image and add up the gradients along the way to the real one [17]. All three take more computing power than Grad-CAM, and that is one reason Grad-CAM is still the one people reach for A fair review has to say that explanations are not proof. One set of sanity checks showed that some saliency methods draw nearly the same picture whether the network has been trained or has random weights [18]; in those tests plain Grad-CAM held up better than its guided variants. Others argue that for individual patients, heatmaps can mislead, and that clinical benefit has to be tested directly [19]. A more radical view is that high-stakes decisions should use models that are interpretable by design [20]. We take a modest position. A heatmap is a useful check for spotting obvious shortcuts, and it should not be read as a guarantee.

VI. UNCERTAINTY ESTIMATION

It helps to separate two kinds of uncertainty [21]. Aleatoric uncertainty comes from the data, such as a blurry film or a finding that experts themselves disagree about. Epistemic uncertainty comes from the model, and it is high when the input is unlike anything seen in training. Dataset shift mostly raises the second kind, which is the kind we care about here.

The easiest confidence score is the sigmoid output. Modern networks tend to be overconfident, i.e., outputs of 0.9 are less than 90% correct [22]. Calibration measures this by sorting predictions into B bins, and comparing the average confidence in each bin to the fraction that was actually correct. Given nb of the N predictions in bin b, the expected calibration error is given by

$$\mathrm{ECE} = \Sigma_b\ (n_b/N)\ |\ \mathrm{acc}(b) - \mathrm{conf}(b)\ | \quad (4)$$

Temperature scaling, which divides the logits by one learned number, is a cheap way to improve it on the data it was tuned on [22]. It does not help much when the data change.

Monte Carlo dropout keeps dropout switched on at test time and runs the image through the network T times [23]. The mean of the T outputs is the prediction, and the spread (or the entropy of the mean) is the uncertainty. For a label c, with $p_c^{(t)}$ the output of pass t,

$$\bar{p}_c = (1/T)\ \Sigma_{t=1..T}\ p_c^{(t)} \quad (5)$$

Deep ensembles train several networks from different random starting points, then check how much those networks disagree [24]. They cost more, because every member needs its own training run, but they are usually the more reliable choice. One large benchmark tested uncertainty under dataset shift. It found that quality got worse for every method as the shift grew, and that ensembles got worse the most slowly [25]. Table II sums up the trade-offs.

TABLE II Common ways to estimate confidence

Method	Extra cost	What it captures	Weak spot
Sigmoid output	None	Raw model confidence	Overconfident
Temperature scaling	One parameter	Calibration	Weak under shift
MC dropout	T forward passes	Approximate model uncertainty	Needs dropout; quality varies
Deep ensemble	M trained models	Disagreement between models	M times the training

Once you have an uncertainty score, you can use it for selective prediction. The system answers only when it is confident enough and hands the rest to a human. A risk-coverage curve shows the trade-off between how many cases it answers and how often it gets them wrong [26]. For a screening tool, this is probably the most practical way to use uncertainty.

VII. OUR EVALUATION: THE TRUSTXRAY PROJECT

Our project, TrustXray, puts the three pieces from the earlier sections into one pipeline (Fig. 1). A DenseNet121 starts from ImageNet weights and is trained on NIH ChestX-ray14 for all 14 findings. The 112,120 images are split into 77,994 for training, 8,530 for validation and 25,596 for testing. Per-finding thresholds are chosen on the validation set, and the test set is used only once, to report AUC with confidence intervals, sensitivity, specificity and F1. The saved model is then run on CheXpert, restricted to the seven findings the two datasets share.

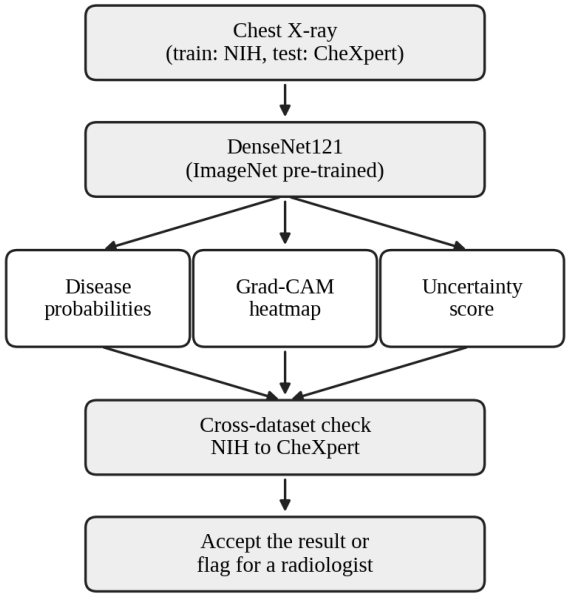

Fig. 1. The TrustXray pipeline: one network, three outputs, and a cross-dataset check before a result is trusted.

We trained two versions on a laptop with an RTX 4050 GPU, at around 250 images per second. The baseline used standard augmentation. The second version, exp1, had a stronger augmentation pipeline in the hope that it would shrink the drop on CheXpert. Training was stopped early if the validation means AUC did not improve for several epochs, and the best checkpoint by validation AUC was retained.

TABLE III Validation results on NIH ChestX-ray14 (mean AUC over 14 findings)

Measure	Baseline	Strong aug. (exp1)
Best validation AUC	0.8353 (ep. 6)	0.8312 (ep. 6)
Validation AUC, epoch 8	0.8175	0.8285
Training loss, epoch 8	0.1036	0.1173

Table III shows what we have so far. Stronger augmentation did not raise the best validation AUC; the two runs are within 0.005 of each other, which on a single run we do not treat as a real difference. What

did change is the behaviour after epoch 6. The baseline's validation AUC fell from 0.8353 to 0.8175 while its training loss kept dropping, which looks like overfitting. In exp1 the validation AUC stayed at about 0.83 and the training loss was higher, so the model was fitting the training set less tightly. Now, let's be clear about what this doesn't show.

These are validation numbers from single runs each. We are still missing the NIH test results and the CheXpert results (the latter being the part that actually answers the generalisation question) and we have not yet added Grad-CAM maps and uncertainty scores to the evaluation. Until then, we cannot know whether more augmentation can reduce the cross-dataset gap. We will report either way, including a null result.

VIII. OPEN PROBLEMS

Several problems keep coming up. Label noise is the first: until labels are checked by radiologists, we cannot separate a model failure from a label failure. The second is the evidence on explanations, since there is still no agreed way to test whether a heatmap is faithful to what the model did, and most chest X-ray papers show a few well-chosen examples.

The third is uncertainty under shift, where even the best methods get worse when they are needed most. Last, nearly all the work is retrospective. A model that looks trustworthy on a benchmark has not yet been shown to help a radiologist at work, and that needs prospective studies, with attention to how performance varies across patient groups.

IX. CONCLUSION

Chest X-ray classifiers are accurate on the data they were trained on, and that is not the same as being safe to use somewhere else. The literature gives us three partial answers: test across datasets, show where the model is looking, and report how sure it is. Each helps, and each has known weaknesses.

What is missing is evaluation that combines them honestly. Our next steps are to finish the NIH and CheXpert comparison, add Grad-CAM and uncertainty outputs to it, and check whether the model abstains more often on the dataset it has not seen.

ACKNOWLEDGMENT

We would like to sincerely thank our guide, Prof. D. A. Kamble, for her steady guidance and encouragement throughout this work. We also thank our project coordinator, Prof. G. P. Joshi, whose suggestions helped us stay focused and whose feedback on our drafts made this review much clearer than it would have been otherwise.

We are grateful to our Principal Dr. Y.R. Kharde and to the Head of the Department Dr. V.V. Patil, along with the faculty of the Department of Artificial Intelligence and Data Science at Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, for giving us the support and resources we needed. We thank the groups who released the NIH ChestX-ray14 and CheXpert datasets for public use, since this work would not have been possible without them. We also thank the authors of the papers we cited, whose research gave us a solid base to build on.

Finally, we thank our friends and families for their patience and support throughout this work.

REFERENCES

[1] X. Wang *et al*., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2097–2106.

[2] J. Irvin *et al*., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 590–597.

[3] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[4] J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLoS Medicine, vol. 15, no. 11, Art. no. e1002683, Nov. 2018.

[5] E. H. P. Pooch, P. Ballester, and R. C. Barros, "Can we trust deep learning-based diagnosis? The impact of domain shift in chest radiograph classification," in Thoracic Image Analysis, 2020, pp. 74–83.

[6] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nature Machine Intelligence, vol. 3, pp. 610–619, 2021.

[7] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, Art. no. 317, 2019.

[8] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," Medical Image Analysis, vol. 66, Art. no. 101797, 2020.

[9] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," Academic Radiology, vol. 27, no. 1, pp. 106–112, 2020.

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.

[11] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction," in Proc. Mach. Learn. Res. (MIDL), vol. 121, 2020, pp. 136–155.

[12] Y. Ganin et al., "Domain-adversarial training of neural networks," Journal of Machine Learning Research, vol. 17, no. 59, pp. 1–35, 2016.

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, Art. no. 60, 2019.

[14] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 4765–4774.

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. Int. Conf. *Mach. Learn. (ICML)*, 2017, pp. 3319–3328.

[18] J. Adebayo et al., "Sanity checks for saliency maps," in Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 9505–9515.

[19] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," The Lancet Digital Health, vol. 3, no. 11, pp. e745–e750, 2021.

[20] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.

[21] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 5574–5584.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.

[23] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050–1059.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 6402–6413.

[25] Y. Ovadia *et al*., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 13991–14002.

[26] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4878–4887.